REVUE DE STATISTIQUE APPLIQUÉE

R. TOMASSONE

Analyse multidimensionnelle et classification

Revue de statistique appliquée, tome 18, n° 4 (1970), p. 29-34

<http://www.numdam.org/item?id=RSA_1970__18_4_29_0>

ANALYSE MULTIDIMENSIONNELLE ET CLASSIFICATION

R. TOMASSONE

INRA, Laboratoire de Biométrie du CNRZ, 78-Jouy en Josas

Les domaines les plus variés font appel à des méthodes de classification : l'archéologie, l'agronomie, la recherche médicale et les sciences économiques pour n'en citer que quelques uns. On classe des objets trouvés lors de campagnes de fouilles, on classe des traitements appliqués à des cultures végétales ou à des espèces animales, on classe les symptomes d'une maladie, et on segmente des processus économiques (le terme de segmentation est plus employé par les économistes que celui de classification ou de taxinomie). Cette diversité des domaines d'application est à la fois passionnante -par les relations interdisciplinaires qu'elle suscite- et dangereuse, chacun croit découvrir des techniques originales et, naturellement, met au point de "nouveaux" programmes pour ordinateur.

Il est donc capital d'étudier ce qui unit les différentes techniques ; c'est-à-dire qu'il faut analyser les concepts les plus importants que nous devrons utiliser au cours des étapes d'une étude de classification.

1 - LES ETAPES D'UNE ETUDE DE CLASSIFICATION.

Une étude classique nous conduira d'une matrice des données de base (un tableau des éléments à classer, sur lesquels nous avons relevé ou mesuré un certain nombre de variables), jusqu'à une classification en groupes des éléments. Généralement les éléments ne pourront appartenir qu'à un groupe à la fois, et nous représenterons la classification par un dendrogramme.

Ces étapes, ainsi que les noms des différentes techniques d'analyse à plusieurs variables, sont schématisés sur la figure ci-après.

2 - DEFINITIONS DES STRUCTURES DANS L'UNIVERS MULTIDIMENSIONNEL.

2.1. La destruction de la structure de référence :

Si nous admettons que la matrice des données de base définit une structure de référence, les transformations que nous lui faisons subir, au cours des différentes étapes de l'analyse, détruisent cette structure. La structure de travail à laquelle on aboutit ne contient pas toute l'information de celle

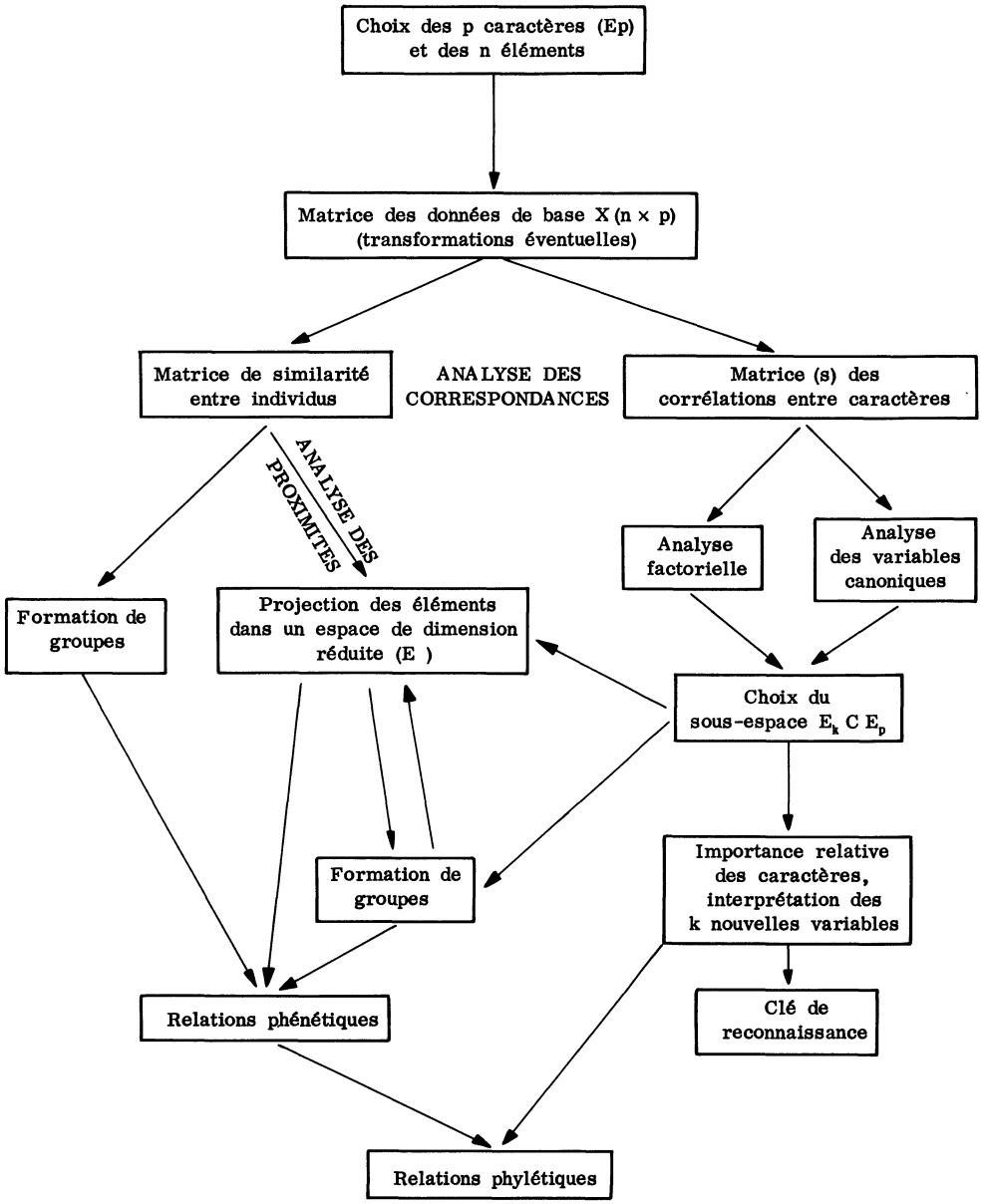

Figure 1 - Les étapes d'une étude de classification (Extrait de MILLIER C. et TOMASSONE R. op. cit.)

de référence ; elle est déformée, et cette déformation n'est pas toujours faite de façon unique. Par exemple, il est bien connu qu'une analyse en composantes principales ne fournit pas les mêmes résultats selon que l'on travaille sur des covariances ou sur des corrélations.

On passe d'une structure à l'autre par le choix d'un modèle ; les différents modèles appartiennent à deux grandes classes :

a) les modèles métriques : ce sont ceux qui font appel aux covariances entre caractères et aux distances statistiques ou non entre les éléments.

b) les modèles non-métriques : qui sont généralement bien adaptés aux méthodes non paramétriques. Les mesures qui en découlent ont un sens soit du point de vue de la théorie des probabilités soit du point de vue de la théorie de l'information.

2.2. Structure dans un espace métrique :

Deux aspects sont particulièrement intéressants :

2.2.1. Dualité des solutions : que l'on analyse d'abord les liens entre variables pour mieux regrouper les éléments, ou que les voisinages entre éléments nous aident à mieux comprendre les "facteurs" sous-jacents aux caractères les résultats sont complémentaires comme l'a montré GOWER J.C. (1966). Cette dualité n'est d'ailleurs pas caractéristique de l'espace métrique puisqu'elle est magnifiquement illustrée par l'analyse des correspondances où variables et éléments jouent des rôles identiques (on peut d'ailleurs les retrouver sur un même graphique).

2.2.2. Fonctions de distances : nous ne voulons parler ici que de deux types de fonctions :

a) celle de Mahalanobis qui nous parait très caractéristique de l'apport très original de l'analyse multidimensionnelle. En plus de ses propriétés statistiques, quelquefois intéressantes, elle intègre deux caractéristiques pour mesurer une distance :

1/ les différences entre toutes les valeurs moyennes des caractères pris en compte.

2/ les liaisons entre ces caractères.

On peut voir facilement qu'une variable apparemment sans intérêt, si elle est étudiée seule, peut se révéler extrèmement discriminante en association avec d'autres.

b) celle de Gower qui permet de mesurer une distance avec plusieurs variables quelle que soit la nature de ces variables (continues, discrètes ordonnées ou non).

3 - ORDINATION ET CLASSIFICATION :

Nous voulons parler ici des différentes stratégies possibles dans l'analyse d'un tableau de données en vue d'une classification.

3.1. Les méthodes d'ordination :

Elles appartiennent à l'analyse multivariable classique, même si dépouillée de son sens statistique, elle n'est utilisée que dans un but descriptif. Elles impliquent une simplification de l'espace E_p à p dimensions -l'espace de référence- en un espace E_k à k dimensions ($k < p$) - l'espace de travail. Ce dernier contient presque toute l'information du premier, il est généralement de dimension plus réduite, il s'analyse toujours plus simplement. On peut distinguer deux grands groupes de méthodes :

a) celui relatif aux différentes analyses factorielles :

- l'analyse en composantes principales où l'accent est mis sur la variabilité entre les éléments.

- l'analyse factorielle classique surtout employée en psychologie ; on essaie ici de retrouver les corrélations entre les variables à l'aide de facteurs inconnus mais qu'il est possible d'estimer à l'aide des observations sur les variables.

- l'analyse factorielle des correspondances, que nous avons déjà citée, qui s'attache à retrouver des "profils" de variables voisins entre les éléments.

- l'analyse de dispersion (ou analyse des variables canoniques) où sont simultanément étudiés deux niveaux de variations : celui propre aux variations entre groupes d'éléments connus a priori et celui propre aux variations à l'intérieur de ces groupes. L'analyse de dispersion peut aussi s'envisager comme l'étude de la liaison entre deux structures de variables : la première est celle des variables effectivement mesurées, la seconde celle des variables définissant l'appartenance des éléments à des groupes.

b) l'analyse des proximités : connaissant les distances entre trois points, il est possible de reconstruire exactement un triangle dont les longueurs des côtés sont ces distances. Connaissant toutes les distances entre les éléments soumis à l'analyse, on peut retrouver un espace dans lequel les positions relatives des éléments déforment le moins possible ces distances. Plus faible sera la dimension de cet espace ; plus grande sera la simplification apportée.

3.2. Les méthodes de classification divisives :

Elles classent d'abord les éléments en deux groupes aussi distincts que possible ; puis elles recommencent l'opération sur les groupes déjà créés. Employées sans précaution, elles sont dangereuses : si le résultat de la première opération ne peut pas être modifié, un élément sera classé une fois pour toutes (dans le cas de l'analyse d'association de Williams et Lambert par exemple). Il est donc souhaitable de compléter ces méthodes par des algorithmes d'échanges qui peuvent faire passer les éléments d'un groupe à un autre afin d'affiner la classification.

3.3. Les méthodes de classification agglomératives :

Elles commencent par regrouper les éléments les plus voisins ; puis les autres sont rattachés à des niveaux plus ou moins éloignés. SNEATH P.H.A. (1968) voit dans les principes qui régissent la formation des groupes trois concepts :

a) celui de masse où chaque nouvel élément qui s'intègre à un groupe en accroit la dimension, la variabilité et le pouvoir d'attraction.

b) celui de densité où les centres des groupes doivent comporter davantage d'éléments que les frontières.

c) celui de réseau où les distances ne sont pas importantes par leur valeur absolue mais par l'ordre qu'elles définissent entre elles. Plus indépendantes de la définition analytique de la fonction de distance, les méthodes qui utilisent ce concept sont souvent sensibles aux effets de chaîne ; il est donc indispensable d'en limiter les conséquences, WISHART, D. (1968). Les deux premiers concepts donnent naissance à des méthodes qui attachent une grande importance à l'homogéinité intragroupe.

A ces concepts qui caractérisent les groupes formés, s'ajoutent des processus distincts dans leur formation ; nous pouvons en distinguer quatre :

a) comment créer les premiers groupes ?

b) comment fusionner des éléments nouveaux à des groupes déjà créés ?

c) comment interdire une nouvelle fusion ?

d) comment échanger des éléments entre des groupes ?

4 - CONCLUSION

Souvent mal formalisées les méthodes de la taxinomie numérique font quelquefois figure de "recettes". Vues d'un très mauvais oeil par nombre de théoriciens, elles répondent à un besoin concret clairement formulé par de nombreux utilisateurs. Il est particulièrement important d'aborder une étude de classification à la fois par des méthodes d'ordination et par des méthodes de classification, les deux s'éclairant mutuellement. Il est, enfin, important de noter que la majorité des méthodes actuelles ne permettent de classer que quelques centaines d'éléments (d'une centaine à un millier selon la taille de l'ordinateur dont on dispose).

5 - BIBLIOGRAPHIE SOMMAIRE. (se reporter à (4) et (6) pour une bibliographie plus abondante)

[1] GOWER J.C. - "Some distance properties of latent root and vector method used in multivariate analysis", Biometrika (1966) 53, 325-338

[2] GOWER J.C. - "A general coefficient of similarity and some of its properties", Roneo Rothamsted Exp. Stat. (1968), 16 p.

[3] LERMAN I.C. - Les bases de la classification automatique Gauthier-Villars, Paris, 1970, 117 p.

[4] MILLIER C. et TOMASSONE R. - "Méthodes d'ordination et de classification : leur efficacité et leur limite". dans Coll. Intern. CNRS sur l'emploi des calculateurs en archéologie, Marseille (1969), 207-228.

[5] SNEATH P.H.A. - "Evaluation of clustering methods" dans Coll. Numerical Taxonomy, St. Andrews (1968), 216-225.

[6] SOKAL R.R. et SNEATH P.H.A. - Principles of numerical taxonomy", Freeman, San Francisco (1963), 359 p.

[7] WISHART D. - "Mode Analysis: a generalisation of nearest neighbour which reduces chaining effects" dans Coll. Numerical Taxonomy, St. Andrews (1968), 233-254.